Evaluating the Scientific Veracity of PhD Theses Written under Supervision of Prof. Dr. Jens Förster

Letty Koopman, Frans J. Oort, Chris A.J. Klaassen

September 11, 2017

Letty Koopman
V.E.C.Koopman@uva.nl

Frans J. Oort
F.J.Oort@uva.nl

Chris A.J. Klaassen
C.A.J.Klaassen@uva.nl

Note that the authors do not represent any departmental or institutional affiliation with this report.

Executive Summary

At the request of the board of the University of Amsterdam we investigated the scientific veracity of the four PhD theses written under supervision of Prof. Dr Jens Förster at the University of Amsterdam. We found no evidence of low veracity of the reported results in three of the four theses. However, one thesis has two chapters that show various peculiarities in their results. These chapters have been published as Gillebaart et al. (2012). In fact, under the model used in this publication to analyze the underlying data, the probability is very small that such extreme results as or more extreme results than the ones in the paper will be obtained. We conclude that Gillebaart et al. (2012) shows strong evidence of low scientific veracity of its results. However, since the underlying data have been collected under the responsibility of Dr Förster, we stress that his coauthors should not be blamed for this, in our opinion. Details of our investigation are presented in this report.

Contents

1 Introduction

Four PhD students have written their PhD thesis at the University of Amsterdam under supervision of Prof. Dr Jens Förster. These theses are reviewed using three methods. Two of these methods have also been used in the report of Peeters et al. (2015), namely the ΔF test for ANOVA models with Fisher's test for combined probabilities and the EV-method with Evidential Values EV in favor of dependence versus independence. Our third method is based on the statistic $Z_{\mathbf{V}}$, which is basic to the EV-method. This third method is presented and explained in Appendix A. All three methods can be applied only if sample averages and sample standard deviations are given for three independent conditions or groups; typically indicated by high, control, low. This situation often occurs in publications of Förster, but much less so in the theses under study here. For one thesis we performed additional statistical analyses to check the reported results with available data.

Prior to our statistical analyses, we obtained information on the circumstances of data collection for each study from the co-supervisors of the PhD students.

The theses can be found at the University of Amsterdam Digital Academic Repository, UvA-DARE, `http://dare.uva.nl/search`

The report and R-script from Peeters et al. (2015) and discussion of its methods and results can be found via `http://www.uva.nl/en/news-events/news/uva-news/content/news/2015/07/update-articles-jens-forster-investigated.html`

2 PhD Thesis of Lottie Bullens

Table 1: *The four publications of Lottie Bullens*

Chapter	Abbreviation	Publication
2	LB2	Bullens, L., van Harreveld, F., & Förster, J. (2011). Journal of Experimental Social Psychology, **47**, 800–805
3	LB3	Bullens, L., van Harreveld, F., Förster, J., & van der Pligt, J. (2013). Journal of Experimental Social Psychology, **49**, 1093–1099
4	LB4	Bullens L., van Harreveld, F., Förster, J., & Higgens, E.T. (2014). Journal of Experimental Psychology: General, **143**, 835–849
5	LB5	Bullens L. & van Harreveld, F. Manuscript under revision

Bullens collected all data for her studies that are reported in her PhD thesis, herself. Her studies have research designs to which our methods to investigate veracity do not apply.

3 PhD Thesis of Laura Dannenberg

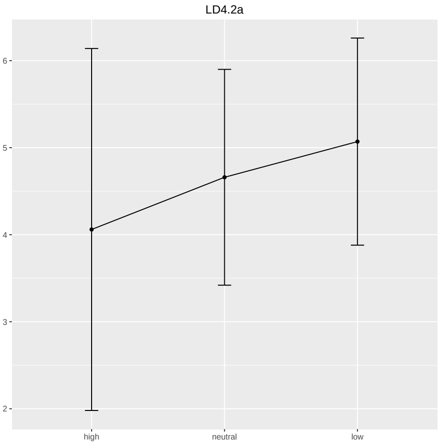
3.1 Overview

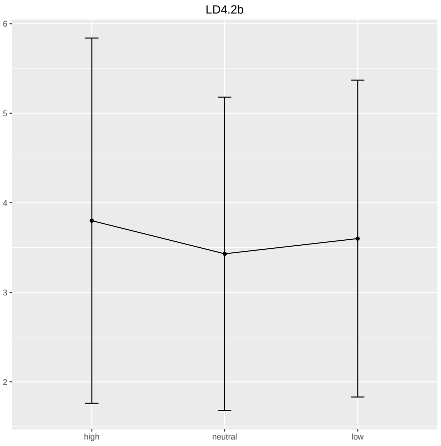
Table 2: *The four publications of Laura Dannenberg*

Chapter	Abbreviation	Publication
2	LD2	Dannenberg, L., Förster, J., & Jostmann, N.B. (2009). Jaarboek Sociale Psychologie 2009, 65–72
3	LD3	Dannenberg, L., Förster, J., & Jostmann, N.B. (2012). Consciousness and Cognition, **21**, 456–463
4	LD4	Dannenberg, L., Jostmann, N.B., & Förster, J. In preparation: Power reduces illusions of agency

Dannenberg collected all data for the studies that are reported in her PhD thesis, herself. In Chapters 2, 3, and part of 4 only two-group designs are used; therefore, our methods to investigate veracity cannot be applied. In Chapter 4 there are three conditions. The results from Chapter 4 do not reflect a consistent difference in means between the conditions. That is, in two subexperiments the *low* condition has a lower mean than the *neutral* condition, but in two other experiments the mean of the *low* condition is higher than the mean of the *neutral* condition. We therefore conducted our analyses of these experiments twice, once with order *high, neutral, low* and once with order *high, low, neutral.*

3.2 Analysis by ΔF and EV





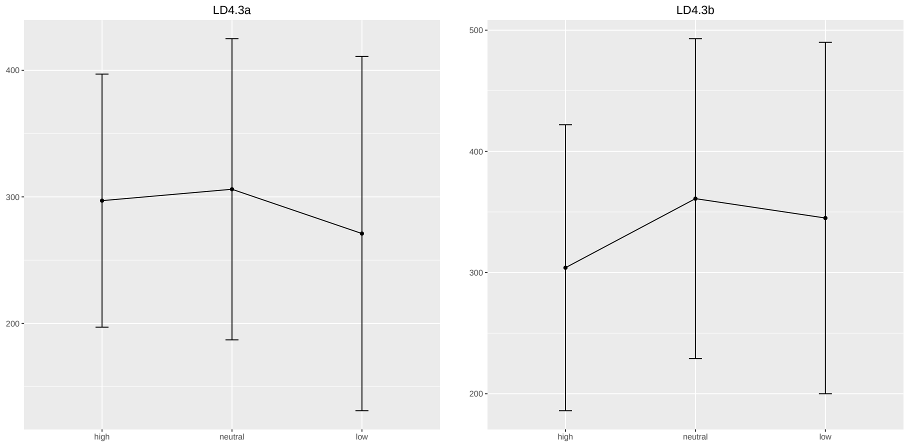


Figure 1. *Trend lines for the means of participant scores for independent groups: high, neutral, low. The error bars represent distances of one standard deviation from the cell mean.*

Table 3: *Results for ΔF, the associated probability $p(\Delta F)$, and evidential value EV for the Dannenberg studies, with group ordering high, neutral, low. n = number of observations per condition, $m_{\text{high/neut/low}}$ = mean of the high/neutral/low condition, $s_{\text{high/neut/low}}$ = standard deviation of the high/neutral/low condition.*

	n	m_{high}	m_{neut}	m_{low}	s_{high}	s_{neut}	s_{low}	ΔF	$p(\Delta F)$	EV
LD4.2a	83/3	4.06	4.66	5.07	2.08	1.24	1.19	0.069	0.793	2.171
LD4.2b	83/3	3.80	3.43	3.60	2.04	1.75	1.77	0.394	0.532	1.156
LD4.3a	56/3	297	306	271	100	119	140	0.420	0.520	1.149
LD4.3b	56/3	304	361	345	118	132	145	0.967	0.330	1

Fisher	Overall EV
0.2695861	2.883618

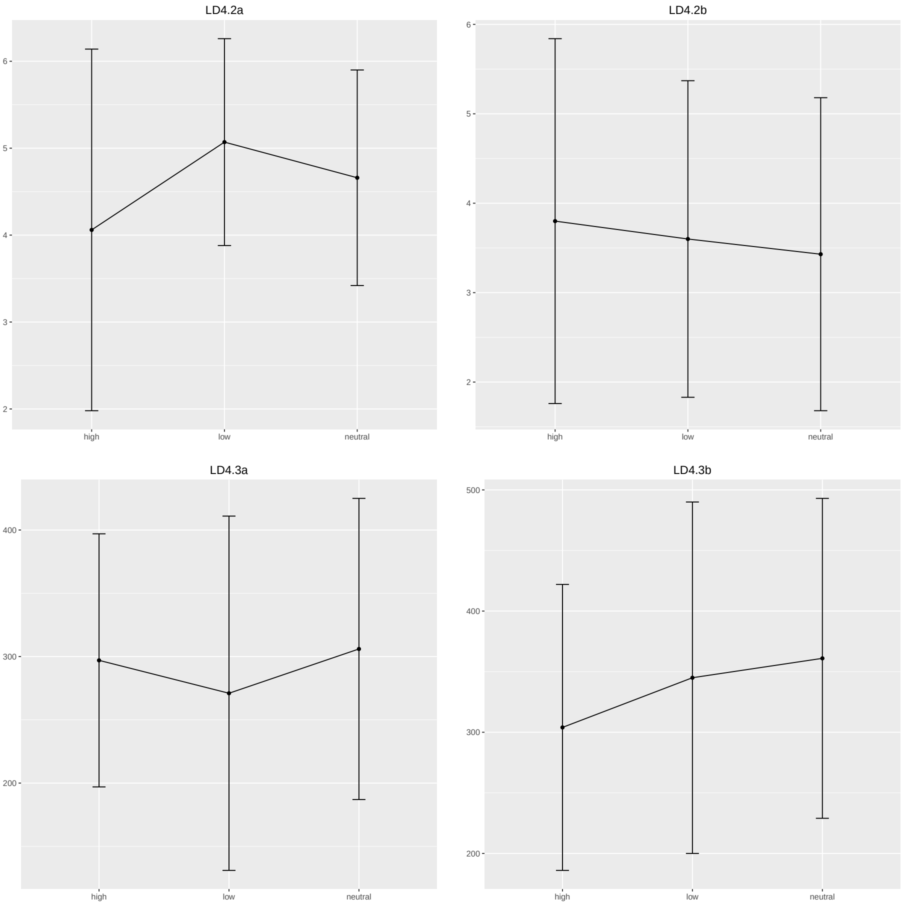


Figure 2. *Trend lines for the means of participant scores for independent groups: high, low, neutral. The error bars represent distances of one standard deviation from the cell mean.*

Table 4: *Results for ΔF, the associated probability $p(\Delta F)$, and evidential value EV for the Dannenberg studies, with group ordering high, low, neutral. n = number of observations per condition, $m_{\text{high/neut/low}}$ = mean of the high/neutral/low condition, $s_{\text{high/neut/low}}$ = standard deviation of the high/neutral/low condition.*

	n	m_{high}	m_{low}	m_{neut}	s_{high}	s_{low}	s_{neut}	ΔF	$p(\Delta F)$	EV	EV_{up}
LD4.2a	83/3	4.06	5.07	4.66	2.08	1.19	1.24	3.878	0.052	1	
LD4.2b	83/3	3.80	3.60	3.43	2.04	1.77	1.75	0.001	0.972	14.542	16.993
LD4.3a	56/3	297	271	306	100	140	119	0.808	0.373	1.031	
LD4.3b	56/3	304	345	361	118	145	132	0.113	0.738	1.991	

Fisher	Overall EV
0.6183159	29.85067

The statistics in Tables 3 and 4 yield *no evidence* of low veracity of the results according to the guidelines in Subsection 1.4 of Peeters et al. (2015). Since the order of means in the conditions is not consistent among the different studies, we conducted our analyses in two ways. Firstly with the condition ordering of *high, neutral, low*, which gives no evidence of low veracity. Secondly with the condition ordering of *high, low, neutral*, where one study, 4.2b, yields an evidential value more than 6, which is deemed substantial according to the guidelines. However, with four studies, the guidelines require at least two substantial evidential values to suggest low veracity of the results. In addition, Fisher's combined probability test yields a left-tail probability of 1 - 0.618 = 0.382, which is not even close to the value of 0.999, which would indicate low veracity of the results. Nevertheless, we also present the results of the method outlined in Appendix A as follows.

3.3 Analysis by $Z_{\mathbf{V}}$

In terms of the notation used in Table 3 the value

$$z_{\mathbf{V}} = \frac{\sqrt{n}(m_{\text{high}} - 2m_{\text{neut}} + m_{\text{low}})}{\sqrt{s^2_{\text{high}} + 4s^2_{\text{neut}} + s^2_{\text{low}}}} \tag{1}$$

can be viewed as a realization of the statistic $Z_{\mathbf{V}}$ from Appendix A. Tables 5 and 6 below present the $|Z_{\mathbf{V}}|$ values for the data from Tables 3 and 4, respectively. In their last column Tables 5 and 6 also show the p-values

Table 5: *Results for $|Z_{\mathbf{V}}|$ and the associated upperbounds on the probabilities $p_1 = P\left(A_{(1)} \leq x,\ A_{(2)} \leq y\right) \leq p_2 = P\left(A_{(1)} \leq A_{(2)} \leq y\right)$ computed via (12) with x the smallest and y the second smallest among the $|Z_{\mathbf{V}}|$ values of the independent studies in Table 3. n = number of observations per condition, $m_{\text{high/neut/low}}$ = mean of the high/neutral/low condition, $s_{\text{high/neut/low}}$ = standard deviation of the high/neutral/low condition.*

	n	m_{high}	m_{neut}	m_{low}	s_{high}	s_{neut}	s_{low}	$\lvert Z_{\mathbf{V}}\rvert$	$p_1 \leq p_2$
LD4.2a	83/3	4.06	4.66	5.07	2.08	1.24	1.19	0.28979	
LD4.2b	83/3	3.80	3.43	3.60	2.04	1.75	1.77	0.64248	
LD4.3a	56/3	297	306	271	100	119	140	0.64732	0.51619
LD4.3b	56/3	304	361	345	118	132	145	0.97498	0.65604

from formulae (18) and (20) of Appendix A. These p-values are computed under the assumption that (12) is valid. Since the neutral priming condition is assumed to be a condition in between the high and low conditions, Table 5 considers the natural order of the priming conditions and it seems reasonable to use (12) and not just the weaker inequality (13). However, for reasons explained in the preceding Subsections we have also added Table 6, again using (12).

Table 6: *Results for $|Z_{\mathbf{V}}|$ and the associated upperbounds on the probabilities $p_1 = P\left(A_{(1)} \leq x,\ A_{(2)} \leq y\right) \leq p_2 = P\left(A_{(1)} \leq A_{(2)} \leq y\right)$ computed via (12) with x the smallest and y the second smallest among the $|Z_{\mathbf{V}}|$ values of the independent studies in Table 4. n = number of observations per condition, $m_{\text{high/low/neut}}$ = mean of the high/low/neutral condition, $s_{\text{high/low/neut}}$ = standard deviation of the high/low/neutral condition.*

	n	m_{high}	m_{low}	m_{neut}	s_{high}	s_{low}	s_{neut}	$\lvert Z_{\mathbf{V}}\rvert$	$p_1 \leq p_2$
LD4.2a	83/3	4.06	5.07	4.66	2.08	1.19	1.24	2.19980	
LD4.2b	83/3	3.80	3.60	3.43	2.04	1.77	1.75	0.03550	
LD4.3a	56/3	297	271	306	100	140	119	0.82295	0.06066
LD4.3b	56/3	304	345	361	118	145	132	0.31789	0.26076

Table 5 nor Table 6 contains indications of low veracity of the results in Chapter 4 of Dannenberg's PhD thesis, which is in line with the conclusion of the preceding Subsection.

4 PhD Thesis of Marleen Gillebaart

4.1 Overview

Table 7: *The two publications of Marleen Gillebaart*

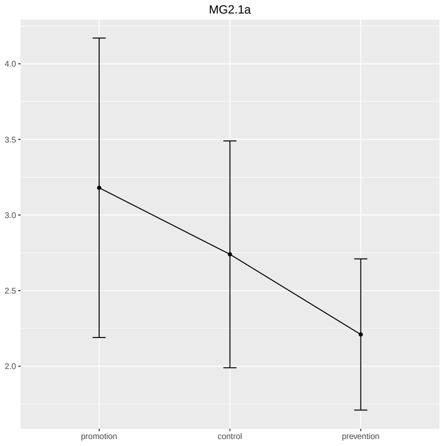
Chapter	Abbreviation	Publication
2	MG2	Gillebaart, M., Förster, J., & Rotteveel, M. (2012). Journal of Experimental Psychology: General, **141**, 699–714
3	MG3	Gillebaart, M., Förster, J., & Rotteveel, M. (2012). Journal of Experimental Psychology: General, **141**, 699–714
4	MG4	Gillebaart, M., Förster, J., Rotteveel, M., & Jehle, A. C. (2013). Creativity Research Journal, **25**(3), 280–285

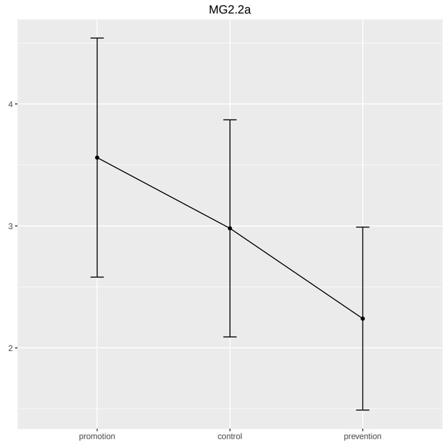
Gillebaart collected the data for the studies that are reported in Chapter 4 of her PhD thesis, herself. Chapters 2 and 3 are based on a single publication reporting six studies conducted in Germany, unbeknown to Gillebaart, without any contribution by Gillebaart. Förster gave Gillebaart processed versions of the six data sets. Gillebaart subsequently conducted statistical analyses of these data sets and reported the results in a paper, co-authored by Förster and co-supervisor M. Rotteveel, which was published as Gillebaart et al. (2012).

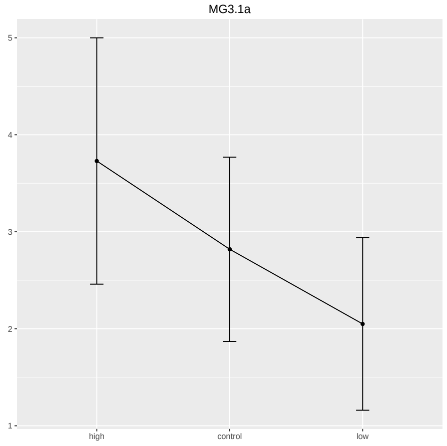
Since the studies reported in Chapters 2 and 3 have been combined into Gillebaart et al. (2012), we will investigate this paper in stead of these chapters. All studies in this publication and in Chapter 4 of the thesis concern three independent groups, enabling us to apply our methods to judge their veracity. However, the experimental designs also include an additional within-subjects factor; each participant in each of the three independent groups is subjected to tests under four or three different conditions. As our veracity analyses are valid for independent measurements only, we conducted the analyses for each of these four or three measurements separately.

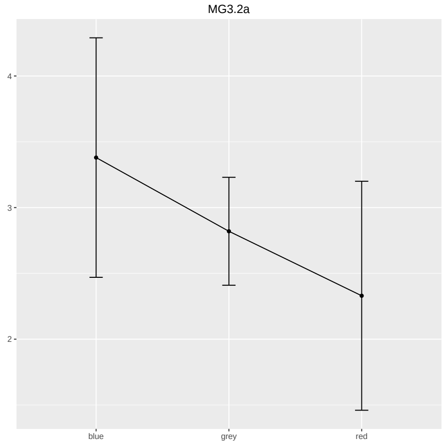
After finding typing mistakes in the published results of the control condition of MG2.1, we used the processed data set of MG2.1 to recompute the means and standard deviations and used these recomputed values in our analyses. We also used the processed data to recompute all F-values that Gillebaart reported, and found that these had been correctly reported.

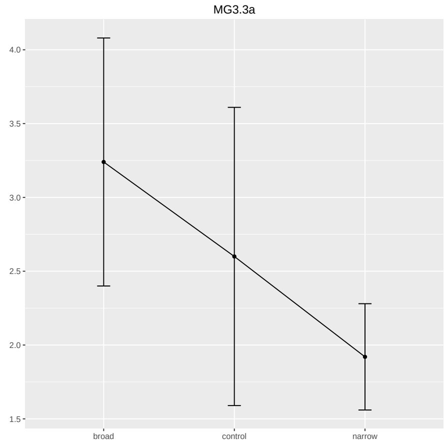
4.2 Analysis by ΔF and EV

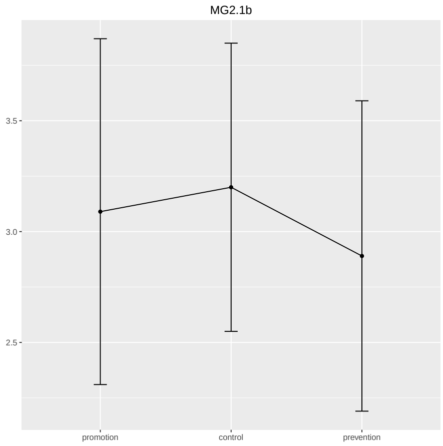


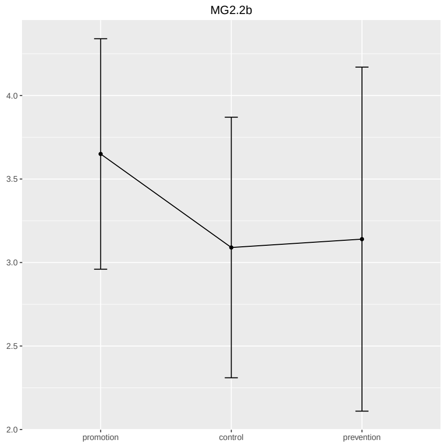


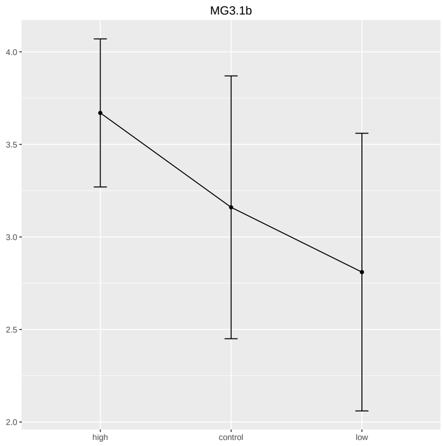


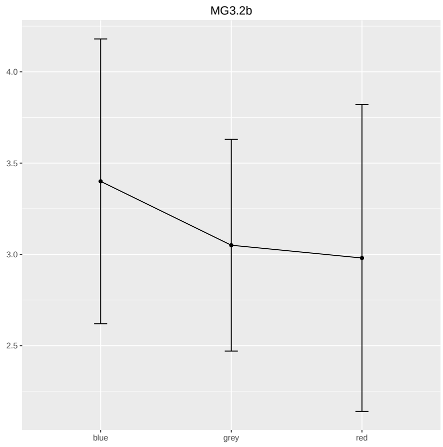


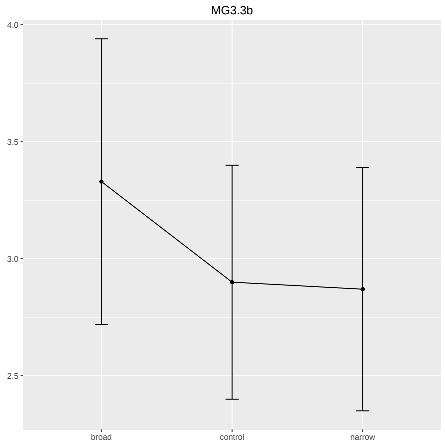


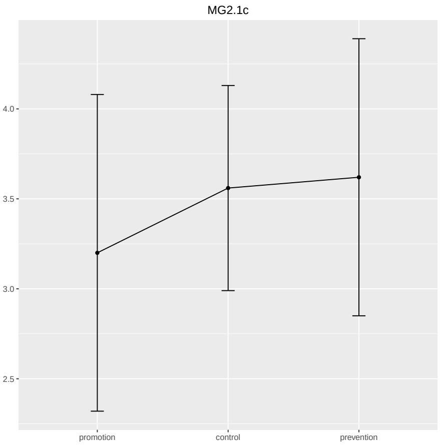

MG2.1b
3.5
3.0
2.5
promotion
control
prevention

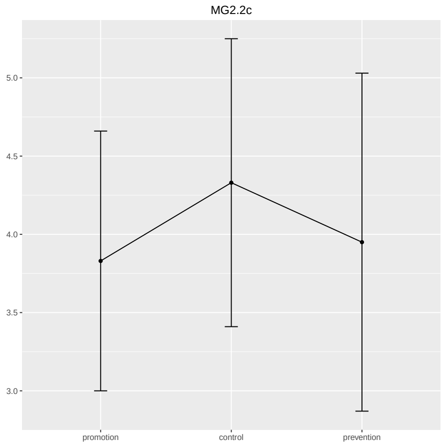

MG2.2b
4.0
3.5
3.0
2.5
2.0
promotion
control
prevention

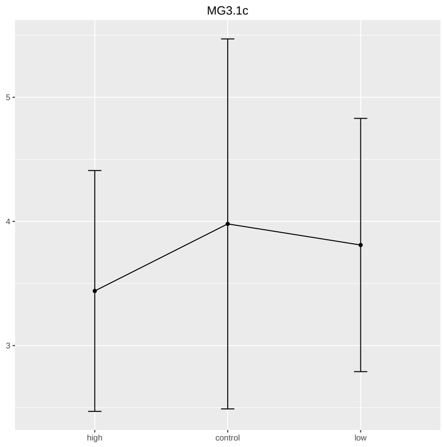

MG3.1b
4.0
3.5
3.0
2.5
2.0
high
control
low

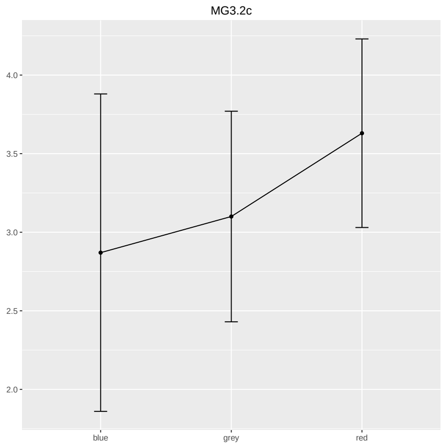

MG3.2b
4.0
3.5
3.0
2.5
blue
grey
red

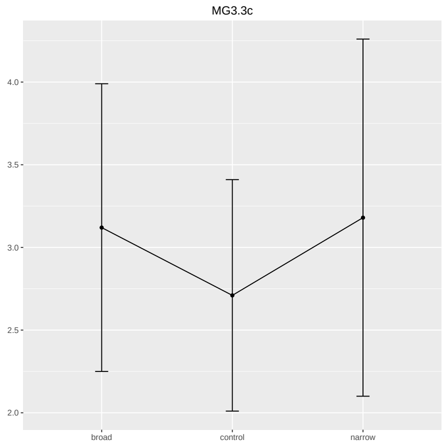

MG3.3b
4.0
3.5
3.0
2.5
broad
control
narrow

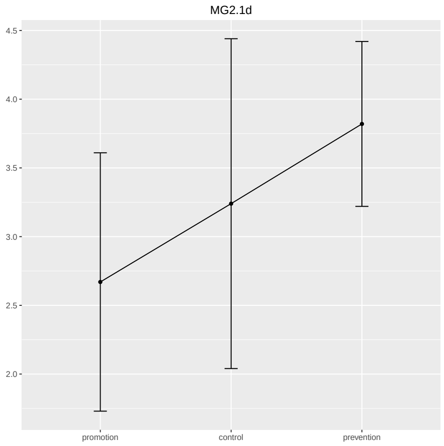

MG2.1c
4.0
3.5
3.0
2.5
promotion
control
prevention

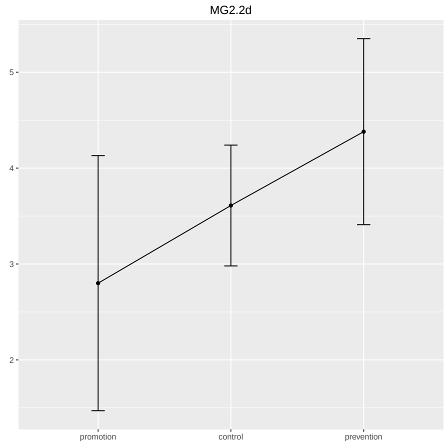

MG2.2c
5.0
4.5
4.0
3.5
3.0
promotion
control
prevention

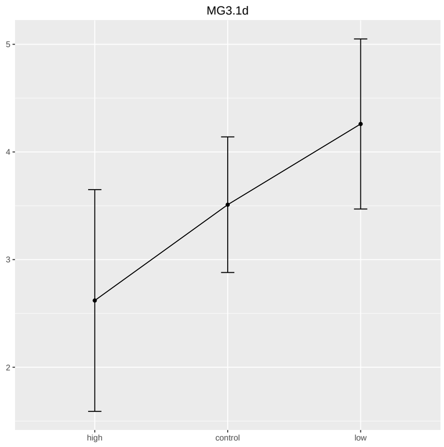

MG3.1c
5
4
3
high
control
low

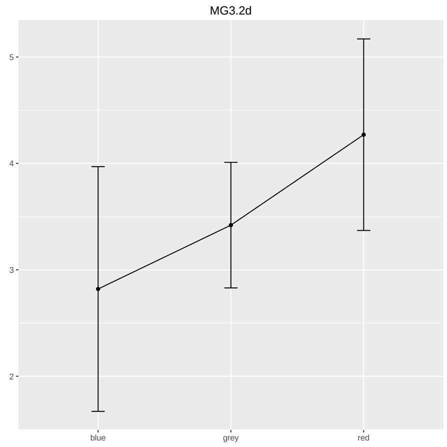

MG3.2c
4.0
3.5
3.0
2.5
2.0
blue
grey
red

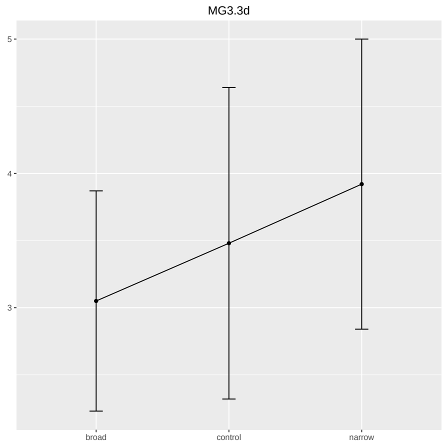

MG3.3c
4.0
3.5
3.0
2.5
2.0
broad
control
narrow


MG2.1d
4.5
4.0
3.5
3.0
2.5
2.0
promotion
control
prevention


MG2.2d
5
4
3
2
promotion
control
prevention


MG3.1d
5
4
3
2
high
control
low


MG3.2d
5
4
3
2
blue
grey
red


MG3.3d
5
4
3
broad
control
narrow

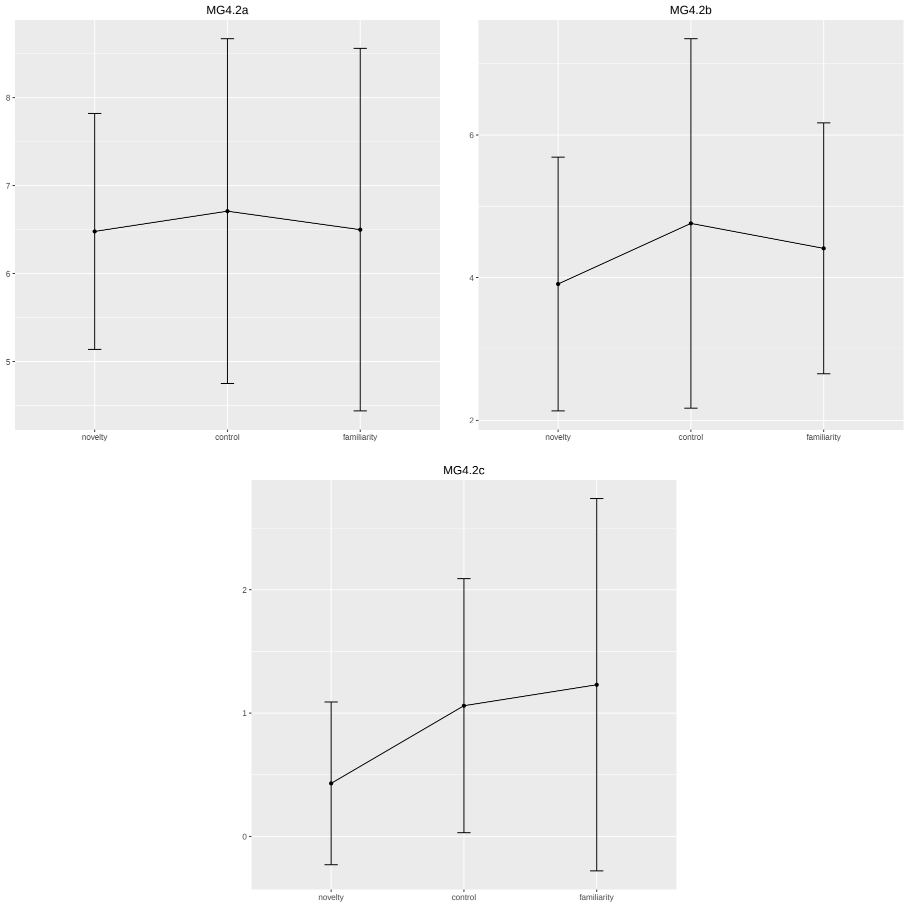


Figure 3. *Trend lines for the means of participant scores arranged into five groups, the first four of which contain independent trend lines. The error bars represent distances of one standard deviation from the cell mean.*

Table 8: *Results for ΔF, the associated probability $p(\Delta F)$, and evidential value EV for the Gillebaart studies. n = number of observations per condition, $m_{\text{left/con/right}}$ = mean of the left/control/right condition, $s_{\text{left/con/right}}$ = standard deviation of the left/control/right condition.*

	n	m_{left}	m_{con}	m_{right}	s_{left}	s_{con}	s_{right}	ΔF	$p(\Delta F)$	EV	EV_{up}
MG2.1a	66/3	3.18	2.74	2.21	0.99	0.75	0.50	0.050	0.82430	2.750	
MG2.2a	58/3	3.56	2.98	2.24	0.98	0.89	0.75	0.111	0.74070	1.939	
MG3.1a	44/3	3.73	2.82	2.05	1.27	0.95	0.89	0.044	0.83405	2.811	
MG3.2a	60/3	3.38	2.82	2.33	0.91	0.41	0.87	0.028	0.86782	2.975	
MG3.3a	41/3	3.24	2.60	1.92	0.84	1.01	0.36	0.006	0.93847	2.665	9.005
MG2.1b	66/3	3.09	3.20	2.89	0.78	0.65	0.70	1.276	0.26296	1	
MG2.2b	58/3	3.65	3.09	3.14	0.69	0.78	1.03	1.734	0.19312	1	
MG3.1b	44/3	3.67	3.16	2.81	0.40	0.71	0.75	0.157	0.69437	1.737	
MG3.2b	60/3	3.40	3.05	2.98	0.78	0.58	0.84	0.475	0.49348	1.061	
MG3.3b	41/3	3.33	2.90	2.87	0.61	0.50	0.52	1.255	0.26947	1	
MG2.1c	66/3	3.20	3.56	3.62	0.88	0.57	0.77	0.585	0.44720	1.02	
MG2.2c	58/3	3.83	4.33	3.95	0.83	0.92	1.08	2.866	0.09591	1	
MG3.1c	44/3	3.44	3.98	3.81	0.97	1.49	1.02	0.900	0.34824	1.03	
MG3.2c	60/3	2.87	3.10	3.63	1.01	0.67	0.60	0.492	0.48586	1.07	
MG3.3c	41/3	3.12	2.71	3.18	0.87	0.70	1.08	2.246	0.14199	1	
MG2.1d	66/3	2.67	3.24	3.82	0.94	1.20	0.60	0	0.98391	3.073	34.227
MG2.2d	58/3	2.80	3.61	4.38	1.33	0.63	0.97	0.005	0.94304	4.840	7.055
MG3.1d	44/3	2.62	3.51	4.26	1.03	0.63	0.79	0.071	0.79175	2.117	
MG3.2d	60/3	2.82	3.42	4.27	1.15	0.59	0.90	0.252	0.61764	1.216	
MG3.3d	41/3	3.05	3.48	3.92	0.82	1.16	1.08	0	0.98825	6.373	43.565
MG4.2a	54/3	6.48	6.71	6.50	1.34	1.96	2.06	0.176	0.67630	1.631	
MG4.2b	54/3	3.91	4.76	4.41	1.78	2.59	1.76	0.999	0.32229	1.014	
MG4.2c	54/3	0.43	1.06	1.23	0.66	1.03	1.51	0.504	0.48087	1.078	

	Fisher	Overall EV
MG Chapters 2 and 3	0.0733815	60075.61
MG Chapter 4	0.3921491	1.782833

According to the guidelines from Peeters et al. (2015) the statistics from Table 8 yield *no evidence* for low veracity for Chapters 2, 3, and 4, as only one of the EVs is larger than 6 (MG3.3d) and Fisher's combined probability test gives a left-tail probability of 1 - 0.071 = 0.929. However, there are three EVs in intervals that include the value 6. Moreover, although the left-tail probability does not meet the 0.999 criterion of *strong evidence* for low veracity, its value is very close to the threshold. These facts have urged us to have a closer look at the sample means and standard deviations in Table 8 and to perform further analyses on Gillebaart et al. (2012).

4.3 Analysis by $Z_{\mathbf{V}}$

In terms of the notation used in Table 8 the value

$$z_{\mathbf{V}} = \frac{\sqrt{n}(m_{\text{left}} - 2m_{\text{con}} + m_{\text{right}})}{\sqrt{s^2_{\text{left}} + 4s^2_{\text{con}} + s^2_{\text{right}}}} \tag{2}$$

can be viewed as a realization of the statistic $Z_{\mathbf{V}}$ from Appendix A. There it is explained that under the model used by Gillebaart et al. (2012) to analyze their data, the probability is very small that $Z_{\mathbf{V}}$ takes on values as close to 0 as or closer to 0 than the ones obtained via (2) from Table 8. The p-values from formulae (18) and (20) of Appendix A are shown in Table 9 below, which contains the same studies as Table 8, but with the studies from Gillebaart et al. (2013) left out.

Table 9: *Results for $|Z_{\mathbf{V}}|$ and the associated upperbounds on the probabilities $p_1 = P\left(A_{(1)} \leq x,\ A_{(2)} \leq y\right) \leq p_2 = P\left(A_{(1)} \leq A_{(2)} \leq y\right)$ computed via (12) with x the smallest and y the second smallest among the $|Z_{\mathbf{V}}|$ values within the indicated group of independent studies. n = number of observations per condition, $m_{\text{left/con/right}}$ = mean of the left/control/right condition, $s_{\text{left/con/right}}$ = standard deviation of the left/control/right condition.*

	n	m_{left}	m_{con}	m_{right}	s_{left}	s_{con}	s_{right}	$\lvert Z_{\mathbf{V}}\rvert$	$p_1 \leq p_2$
MG2.1a	66/3	3.18	2.74	2.21	0.99	0.75	0.50	0.22629	
MG2.2a	58/3	3.56	2.98	2.24	0.98	0.89	0.75	0.32481	
MG3.1a	44/3	3.73	2.82	2.05	1.27	0.95	0.89	0.21861	
MG3.2a	60/3	3.38	2.82	2.33	0.91	0.41	0.87	0.20836	0.10985
MG3.3a	41/3	3.24	2.60	1.92	0.84	1.01	0.36	0.06670	0.19314
MG2.1b	66/3	3.09	3.20	2.89	0.78	0.65	0.70	1.17973	
MG2.2b	58/3	3.65	3.09	3.14	0.69	0.78	1.03	1.34603	
MG3.1b	44/3	3.67	3.16	2.81	0.40	0.71	0.75	0.37025	
MG3.2b	60/3	3.40	3.05	2.98	0.78	0.58	0.84	0.76783	0.76264
MG3.3b	41/3	3.33	2.90	2.87	0.61	0.50	0.52	1.15382	0.87608
MG2.1c	66/3	3.20	3.56	3.62	0.88	0.57	0.77	0.86165	
MG2.2c	58/3	3.83	4.33	3.95	0.83	0.92	1.08	1.69018	
MG3.1c	44/3	3.44	3.98	3.81	0.97	1.49	1.02	0.82504	
MG3.2c	60/3	2.87	3.10	3.63	1.01	0.67	0.60	0.75286	0.90423
MG3.3c	41/3	3.12	2.71	3.18	0.87	0.70	1.08	1.65087	0.90558
MG2.1d	66/3	2.67	3.24	3.82	0.94	1.20	0.60	0.01772	
MG2.2d	58/3	2.80	3.61	4.38	1.33	0.63	0.97	0.08484	
MG3.1d	44/3	2.62	3.51	4.26	1.03	0.63	0.79	0.29638	
MG3.2d	60/3	2.82	3.42	4.27	1.15	0.59	0.90	0.70987	0.00185
MG3.3d	41/3	3.05	3.48	3.92	0.82	1.16	1.08	0.01376	0.00194

Note that in all studies from Table 9 the control group was implicitly assumed to show behavior that is in between the behavior of the other two groups. Therefore, we have applied inequality (12) in the p_1- and p_2-values reported in Table 9. If all studies in Table 9 would have been independent, then

$$\begin{aligned} 0.03059 = p_1 = P\left(A_{(1)} \leq x,\ A_{(2)} \leq y\right) & \qquad (3) \\ \leq P\left(A_{(1)} \leq A_{(2)} \leq y\right) = p_2 = 0.03209 & \end{aligned}$$

would hold with x the smallest and y the second smallest among all $|Z_{\mathbf{V}}|$ values in the table. As these p-values are well below 0.05 and the p-values for the last group in Table 9 are even extremely small, we conclude that serious doubts arise about the way the data for Gillebaart et al. (2012) have been collected and that hence the scientific value of the results in this paper is very low. In other words, there is *strong evidence of low veracity* of these results.

4.4 Exploratory analyses

Noting that we have concluded there is strong evidence of low veracity of the results in Gillebaart et al. (2012) (henceforth GFR), that we have found some errors in this publication, that the six sets of data on which this publication is based were collected and pre-processed outside the awareness of Gillebaart and Rotteveel, and that the analyses were based on these pre-processed data, we decided to further investigate the six (pre-processed) data sets (i.e. the first five sets from Tables 8 and 9 and the additional data from experiment MG2.3, which compares two conditions). A short report is included as Appendix B.

We first recalculated all means and standard deviations of the observed scores, and we conducted the same repeated measures ANOVAs that were reported by GFR. Post hoc power calculations yielded extreme results, due to extremely large effect sizes. To get a better idea of the sizes of the effects reported by GFR, we calculated effect size d (as defined by Cohen (1988)) to express between and within group differences. The resulting effect sizes seem extremely large, especially in view of the subtle experimental differences between and within groups. When we estimated the variances of the effect sizes we unexpectedly encountered negative correlations between repeated measures. We therefore checked the reliability of the observed scores and found that only 8 out of 68 reliability coefficients met the criterion of 0.7 reliability, whereas 24 out of 68 reliability coefficients were zero. We also found negative correlations between items. This finding led us to investigate the interchangeability of the items, because, according to GFR, the items are "valence neutral" and "assignment of the letters to the exposure frequencies was fully counterbalanced over all participants, and order of stimuli was randomized for each participant" (p. 701). However, our results show that the items are not interchangeable. Apparently the items are not valence neutral, nor counterbalanced, nor randomized.

Appendix B gives an overview of the results of all exploratory analyses. On the basis of these results we conclude that either the experiments have not been conducted in the way described by GFR, or errors have been made in the processing of the original data.

4.5 Discussion

To summarize our findings, several inconsistencies and unexpected phenomena are found that cast doubt on the validity of the results in Gillebaart et al. (2012). First, there is strong evidence of low veracity of the results, as shown in Subsection 5.3. Second, some errors are found in the means, standard deviations, and effect sizes that are reported. Third, the post-hoc power of the tests performed is extreme. All values are above 0.96, which seems highly unlikely with sample sizes ranging from 14 to 22 participants per group. Relatedly, the effect sizes are extremely large, much larger than effect sizes that are commonly found in the social sciences in general, or in other studies of the mere exposure effect in particular (according to a meta-analysis of Bornstein (1989)). Fourth, the reliability of the outcome measure is not sufficient to find effects of the independent variables. Fifth, according to the design of the experiments described by Gillebaart et al. (2012), with interchangeable items, the item scores should have equal means, variances, and covariances within combinations of measurements and conditions, but in too many instances the hypothesis of invariance has to be rejected. We therefore conclude that the data do not match with the design of the experiments.

As there does not seem to be any possible substantive explanation for the combination of peculiarities in the data, we suspect that mistakes have been made in the (pre-)processing of the data.

5 PhD Thesis Janina Marguc

5.1 Overview

Table 10: *The three publications of Janina Marguc.*

Chapter	Abbreviation	Publication
2	JM2	Marguc, J., Förster, J., & Van Kleef, G.A. (2011). Journal of Personality and Social Psychology, **101**(5), 883–901
3	JM3	Marguc, J., Van Kleef, G.A., & Förster, J. (2012). Social Psychological and Personality Science, **3**(3), 379–386
4	JM4	Marguc, J., Van Kleef, G.A., & Förster, J. (2015). Creativity and Innovation Management, **24**(2), 207–216

Marguc collected all data for the studies that are reported in her PhD thesis, herself. All but one of her studies have research designs to which our methods to investigate veracity do not apply. In Study 3b of the article on which Chapter 2 is based, Marguc applied a three-group design with two repeated measures, of which one was a global and one a local manipulation. We performed the veracity analysis on both measures, based on the means and standard deviations that we obtained from the processed data set that was handed over by Marguc.

5.2 Analysis by ΔF and EV

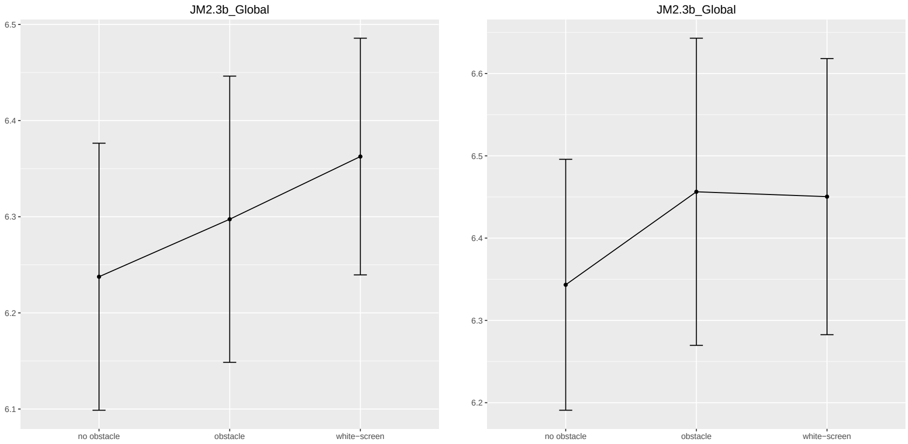


Figure 4. *Trend lines for the means of participant scores for independent groups. The error bars represent distances of one standard deviation from the cell mean.*

Table 11: *Results for ΔF, the associated probability $p(\Delta F)$, and evidential value EV for the Marguc studies. n = number of observations per condition, $m_{\text{ob/nob/white}}$ = mean of the obstacle/no obstacle/white screen condition, $s_{\text{ob/nob/white}}$ = standard deviation of the obstacle/no obstacle/white screen condition.*

	n	m_{nob}	m_{ob}	m_{white}	s_{nob}	s_{ob}	s_{white}	ΔF	$p(\Delta F)$	EV	EV_{up}
JM2.3bGlob	58/3	6.2376	6.2974	6.3626	0.13886	0.14883	0.12306	0.005	0.944	7.889	8.981
JM2.3bLoc	58/3	6.3433	6.4563	6.4504	0.15248	0.18662	0.16777	1.585	0.213	1	

	Fisher	Overall EV
LD chapter 4	0.47639	7.889

The statistics in Table 11 yield *no evidence* of low veracity of the results according to the guidelines of Peeters et al. (2015). Although there is one substantial evidential value, strong evidence of low veracity requires both of the studies to yield a substantial evidential value. In addition, Fisher's combined probability test yields a left-tail probability of 1 - 0.476 = 0.524, not even close to the value of 0.999, which would have indicated low veracity of the results.

5.3 Analysis by $Z_{\mathbf{V}}$

In terms of the notation used in Table 11 the value

$$z_{\mathbf{V}} = \frac{\sqrt{n}(m_{\text{nob}} - 2m_{\text{ob}} + m_{\text{white}})}{\sqrt{s^2_{\text{nob}} + 4s^2_{\text{ob}} + s^2_{\text{white}}}} \tag{4}$$

can be viewed as a realization of the statistic $Z_{\mathbf{V}}$ from Appendix A. Table 12 below presents the $|Z_{\mathbf{V}}|$ values for the data from Table 11. In its last column Table 12 also shows the p-values from formulae (18) and (20) of Appendix A. These p-values are computed under the assumption that (12) is valid. Since the obstacle condition is assumed to be a condition in between the no obstacle and white screen conditions, Table 12 considers the natural order of the priming conditions and it seems reasonable to use (12) and not just the weaker inequality (13).

Table 12 contains no indications of low veracity of the results in Chapter 2 of Marguc's PhD thesis, which is in line with the conclusion of the preceding Subsection.

Table 12: *Results for* $|Z_{\mathbf{V}}|$ *and the associated upperbounds on the probabilities* $p_1 = P\left(A_{(1)} \leq x,\ A_{(2)} \leq y\right) \leq p_2 = P\left(A_{(1)} \leq A_{(2)} \leq y\right)$ *computed via (12) with* x *the smallest and* y *the largest of the two* $|Z_{\mathbf{V}}|$ *values of the independent studies in Table 3.* n *= number of observations per condition,* $m_{\text{high/neut/low}}$ *= mean of the high/neutral/low condition,* $s_{\text{high/neut/low}}$ *= standard deviation of the high/neutral/low condition.*

	n	m_{ob}	m_{nob}	m_{white}	s_{ob}	s_{nob}	s_{white}	$\lvert Z_{\mathbf{V}}\rvert$	$p_1 \leq p_2$
JM2.3bGlob	58/3	6.2376	6.2974	6.3626	0.13886	0.14883	0.12306	0.06769	0.08007
JM2.3bLoc	58/3	6.3433	6.4563	6.4504	0.15248	0.18662	0.16777	1.19717	0.59099

6 Summary

Bullens The publications of Bullens could not be analyzed with our methods to investigate veracity, because the methods do not apply to the two-group research designs of Bullens' studies.

Dannenberg Most of the studies of Dannenberg have research designs to which our methods to investigate veracity do not apply. The other studies show *no evidence* of low veracity.

Gillebaart Chapters 2 and 3 of the PhD thesis of Gillebaart are based on a single article. The studies that are reported in this article show *strong evidence* of low veracity. This conclusion is based on an analysis via the statistic $Z_{\mathbf{V}}$, which is introduced and studied in Appendix A. In addition, other remarkable, even extreme peculiarities in the data are found.

Marguc Most of the studies of Marguc have research designs to which our methods to investigate veracity do not apply. The one study that applied a three-group design shows *no evidence* of low veracity.

A The Statistic $Z_{\mathbf{V}}$

The present and a preceding report about the scientific reliability of publications of Jens Förster have used an ANOVA approach and an Evidential Value approach for analyzing the published results and have received several comments that sometimes show some signs of misunderstanding. This Appendix studies the underlying statistics to these approaches and presents results that are even more convincing than those obtained by the ANOVA and Evidential Value approaches.

A.1 Model and Mathematical Statistical Basis

Consider three sample averages $\bar{X}_{n,1}, \bar{X}_{n,2}$, and $\bar{X}_{n,3}$ of bounded random variables and the corresponding sample standard deviations $S_{n,1}, S_{n,2}$, and $S_{n,3}$, respectively. Note that we assume the three sample sizes to be equal with value n. According to the Central Limit Theorem and the Law of Large Numbers we have

$$\sqrt{n}\left(\bar{X}_{n,i} - \mu_i\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma_i^2\right), \quad S_{n,i} \xrightarrow[n\to\infty]{p} \sigma_i, \quad i = 1, 2, 3, \tag{5}$$

where μ_1, μ_2, μ_3 and $\sigma_1, \sigma_2, \sigma_3$ are the population means and population standard deviations, respectively. This means that for large values of n the sample standard deviation tends to be close to the population standard deviation and that $\bar{X}_{n,i}$ is approximately normal with mean μ_i and variance σ_i^2/n. If the three samples are independent this implies that

$$\bar{X}_{n,1} - 2\bar{X}_{n,2} + \bar{X}_{n,3} \tag{6}$$

is approximately normal with mean $\mu_1 - 2\mu_2 + \mu_3$ and variance $(\sigma_1^2 + 4\sigma_2^2 + \sigma_3^2)/n$. Note that $\sigma_1^2 + 4\sigma_2^2 + \sigma_3^2$ may be estimated by $S_1^2 + 4S_2^2 + S_3^2$ in view of (5).

In some research papers the samples correspond to three groups that get different treatments with the second group being the control group, which typically shows behavior in between the behavior of groups 1 and 3. In an extreme instance of this the population means might satisfy

$$\mu_2 = \frac{\mu_1 + \mu_3}{2}, \tag{7}$$

which is equivalent to

$$\mu_1 - 2\mu_2 + \mu_3 = 0. \tag{8}$$

Of course, it need not be the control group that shows behavior in between the behavior of the other two groups.

If this linearity from (7) and (8) holds, then

$$Z_{\mathbf{V}} = \frac{\sqrt{n}(\bar{X}_{n,1} - 2\bar{X}_{n,2} + \bar{X}_{n,3})}{\sqrt{S_1^2 + 4S_2^2 + S_3^2}} \tag{9}$$

is approximately standard normal.

The problem with quite some experiments/studies in some of the papers of Jens Förster is that this statistic $Z_{\mathbf{V}}$ takes on values pretty close to 0; an event that happens with a pretty small probability for data that are produced according to the model described above. Actually, this statistic $Z_{\mathbf{V}}$ is interpreted in Klaassen (2015) via the forensic statistical approach of evidential values.

In ANOVA studies it is assumed that the variances of all observations are the same, which in our situation means $\sigma_1 = \sigma_2 = \sigma_3$. Under this additional assumption it is natural to replace the statistic $Z_{\mathbf{V}}$ by

$$Z_{\mathbf{C}} = \frac{\sqrt{n}(\bar{X}_{n,1} - 2\bar{X}_{n,2} + \bar{X}_{n,3})}{\sqrt{2\left(S_1^2 + S_2^2 + S_3^2\right)}}, \tag{10}$$

which is also approximately standard normal (provided $\sigma_1 = \sigma_2 = \sigma_3$ holds). The ANOVA approach for the publications of Jens Förster originates from complainant, is based on the statistic $Z_{\mathbf{C}}$, and is also used in our reports.

A.2 The Heart of the Matter

In stead of discussing the evidential value and ANOVA approaches from our reports once more, let's focus here on the underlying statistics $Z_{\mathbf{V}}$ and $Z_{\mathbf{C}}$. As the additional assumption $\sigma_1 = \sigma_2 = \sigma_3$ does not seem to be satisfied in all experiments discussed in the papers by Jens Förster, we will discuss $Z_{\mathbf{V}}$ only.

If the experiment would satisfy the linearity assumption (7) and (8) and sample size wouldn't be too small, then we would have to a good approximation (the same good approximation on which the majority of all statistical analyses in social psychology is based)

$$P\left(|Z_{\mathbf{V}}| \leq z\right) = 2\Phi(z) - 1, \quad z \geq 0, \tag{11}$$

with $\Phi(\cdot)$ the standard normal distribution function. In case the linearity assumption doesn't hold, we have

$$P\left(|Z_{\mathbf{V}}| \leq z\right) \leq 2\Phi(z) - 1, \quad z \geq 0. \tag{12}$$

If it is not clear from the theoretical considerations in the research paper under study which one of the three groups will show behavior in between the behavior of the other two groups, it makes sense to compute the values of the three corresponding absolute values of the $Z_{\mathbf{V}}$ statistics and to use the smallest of these three values as the ultimate $|Z_{\mathbf{V}}|$ statistic. In this case a rough estimate yields

$$P\left(|Z_{\mathbf{V}}| \leq z\right) \leq 3\left(2\Phi(z) - 1\right), \quad z \geq 0, \tag{13}$$

in stead of (12). Note that $|\mu_1 - 2\mu_2 + \mu_3| = 0$ implies $|\mu_1 - 2\mu_3 + \mu_2| = |\mu_2 - 2\mu_1 + \mu_3| = 3|\mu_1 - \mu_3|/2$. Consequently, it is intuitively clear that the more the population means μ_i differ, the closer the behavior of the ultimate $Z_{\mathbf{V}}$ statistic from inequality (13) comes to the behavior of the original $Z_{\mathbf{V}}$ statistic from inequality (12). That is, the more the population means μ_i differ, the more accurate the bound from (12) becomes.

Consider m independent experiments of the above type, compute the values of the corresponding statistics $Z_{\mathbf{V}}$, take the absolute values of them, and call them $A_1, \dots, A_m$. We shall denote the order statistics of $A_1, \dots, A_m$ by $A_{(1)} \leq A_{(2)} \leq \dots \leq A_{(m)}$. So, $A_{(1)}$ denotes the smallest among $A_1, \dots, A_m$. The following result will be proved in Subsection A.4.

Lemma
If $A_1, \dots, A_m$ are independent and identically distributed random variables with continuous distribution function $F(\cdot)$, then

$$P\left(A_{(1)} \leq x,\ A_{(2)} \leq y\right) = 1 - (1 - F(x))^m - mF(x)\,(1 - F(y))^{m-1} \tag{14}$$

holds for $x \leq y$. Furthermore, if $A_1, \dots, A_m, \tilde{A}_1, \dots, \tilde{A}_m$ are independent and identically distributed random variables with distribution function $F(\cdot)$, and $\tilde{A}_{(1)} \leq \tilde{A}_{(2)} \leq \dots \leq \tilde{A}_{(m)}$ are the order statistics of $\tilde{A}_1, \dots, \tilde{A}_m$, then

$$P\left(A_{(1)} \leq \tilde{A}_{(1)},\ A_{(2)} \leq \tilde{A}_{(2)}\right) = \frac{3m - 2}{4(2m - 1)} \tag{15}$$

holds, and

$$\frac{1}{3} \leq \frac{3m - 2}{4(2m - 1)} \leq \frac{3}{8} \tag{16}$$

for $m \geq 2$.

Let us consider now the data from the five independent studies MG2.1d, MG2.2d, MG3.1d, MG3.2d, and MG3.3d from Table 9 from Subsection 4.3. Some numerical computations yield Table 13 below. In the notation of the Lemma this Table shows $A_{(1)} = 0.013757$ and $A_{(2)} = 0.017724$. With the help of the Lemma with $F(x) = 2\Phi(x) - 1,\ x \geq 0$, and $m = 5$ we obtain

$$P\left(A_{(1)} \leq 0.013757,\ A_{(2)} \leq 0.017724\right) = 0.00185. \tag{17}$$

This means the following. Given the model under which the data from these studies have been analyzed in Gillebaart et al. (2012), even without the assumption that the population variances are the same, and given the independence of the five studies, the probability equals 0.00185 that the smallest and second smallest absolute values of the five corresponding $Z_{\mathbf{V}}$-statistics

Table 13: $Z_{\mathbf{V}}$ *values for studies MG2.1d, MG2.2d, MG3.1d, MG3.2d, and MG3.3d in Table 9 from Section 4.3. The (lower bounds to the) evidential values are given as well.*

Study	n	$\bar{X}_{n,1}$ $\bar{X}_{n,2}$ $\bar{X}_{n,3}$	S_1 S_2 S_3	$\lvert Z_{\mathbf{V}}\rvert$	$\mathbf{V}$
MG2.1d	66/3	2.67 3.24 3.82	0.94 1.20 0.60	0.017724	3.073
MG2.2d	58/3	2.80 3.61 4.38	1.33 0.63 0.97	0.084842	4.840
MG3.1d	44/3	2.62 3.51 4.26	1.03 0.63 0.79	0.296379	2.117
MG3.2d	60/3	2.82 3.42 4.27	1.15 0.59 0.90	0.709866	1.216
MG3.3d	41/3	3.05 3.48 3.92	0.82 1.16 1.08	0.013757	6.373

are smaller than the observed values 0.013757 and 0.017724, respectively. If the additional assumption of linearity in the population means is not satisfied, this probability is even smaller. So, we have

$$P\left(A_{(1)} \leq 0.013757,\ A_{(2)} \leq 0.017724\right) \leq 0.00185. \tag{18}$$

To interpret this small probability it should be compared to the average value of this probability when data are collected according to the model. This average is given by (15) with $m = 5$ and hence equals $13/36 \approx 0.36111$. As the probability 0.00185 is considerably smaller than 0.36111, we conclude that the outcomes in Table 13 and hence in the bottom group of Table 9 in Subsection 4.3 are rather unlikely under the model used to analyze these data in Gillebaart et al. (2012).

In contrast, if we would apply the guidelines from Section 1.4 of Peeters et al. (2015) to the studies in Table 13 and the corresponding part of Table 8 from Subsection 4.2, we would have to conclude that there is no evidence for low scientific veracity in these studies, since only one of the evidential values exceeds the value 6. This shows that these guidelines from Peeters et al. (2015) are quite lenient. This is caused by the introduction of the threshold with value 6 thus coarsening the evidential values, which are coarsened functions of the corresponding $Z_{\mathbf{V}}$ values themselves, to indicator quantities.

Another way of interpreting (18) is by hypothesis testing. Choosing as null hypothesis that the data are generated according to the model used in Gillebaart et al. (2012), and as alternative hypothesis that this is not the case, we reject the null hypothesis if

$$A_{(1)} \leq A_{(2)} \leq 0.096 \tag{19}$$

holds. By (14) this test has level 0.05 and the p-value corresponding to $A_{(1)} = 0.013757$ and $A_{(2)} = 0.017724$ equals 0.00185 in view of (18). The classic p-value satisfies

$$P\left(A_{(1)} \leq A_{(2)} \leq 0.017724\right) \leq 0.00194, \tag{20}$$

which is still very small.

In almost every row of Table 9 of Subsection 4.3 the shown sample means differ quite a bit. This suggests that also the corresponding population means will show substantial differences. As argued below inequality (13) this implies that the upperbound at the right hand side of (13) is rather large for application to the results in Table 9 of Subsection 4.3. Nevertheless, we apply (13) in stead of (12) to the bottom group of Table 9 and we obtain

$$\begin{aligned} P\left(A_{(1)} \leq 0.013757,\ A_{(2)} \leq 0.017724\right) &\leq 0.01572, \\ P\left(A_{(1)} \leq A_{(2)} \leq 0.017724\right) &\leq 0.016518. \end{aligned} \tag{21}$$

Of course, these bounds are larger than those from (18) and (20), respectively, but still small.

A.3 Notes

The results of 20 studies in Gillebaart et al. (2012) are presented in Table 9 of Subsection 4.3. These studies are not independent. If they would have been independent and if we would have applied the Lemma with (12) to these results, then we would have gotten

$$P\left(A_{(1)} \leq 0.013757,\ A_{(2)} \leq 0.017724\right) = 0.03059 \tag{22}$$

in stead of (18) and

$$P\left(A_{(1)} \leq A_{(2)} \leq 0.017724\right) = 0.03209 \tag{23}$$

in stead of (20). These results are not surprising as it is easy to check that the probability in (14) is strictly increasing in m.

In any case the p-values (18), (20), (22), and (23) are all well below 0.05 and raise doubt on the scientific veracity of the data in Gillebaart et al. (2012). A notorious instance of doubt about the scientific veracity of data can be found in Mendel (1866). A nice review of this topic is given in Fairbanks and Rytting (2001). Table 2 of ibid. lists the 23 values of χ^2 statistics obtained by application to Mendel's data. Some of these χ^2 values are remarkably small. Applying our Lemma with $1 - F(x)$ equal to the 23 probabilities given in Table 2 of ibid. yields (to prove that the lemma may be applied we transform all χ^2 statistics to uniformly distributed random variables)

$$P\left(A_{(1)} \leq 0.013757,\ A_{(2)} \leq 0.017724\right) = 0.3550 \tag{24}$$

and

$$P\left(A_{(1)} \leq A_{(2)} \leq 0.017724\right) = 0.3574. \tag{25}$$

Comparing these values to (22) and (23) we conclude that it is quite understandable that the dispute about the veracity of Mendel's data is a continuing story and that it is quite reasonable to have doubts about the veracity of the data in Gillebaart et al. (2012).

A.4 Proof of Lemma

Let $x \leq y$. First we notice

$$\begin{aligned} & P\left(A_{(1)} \leq x,\ A_{(2)} \leq y\right) + P\left(A_{(1)} \leq x,\ A_{(2)} > y\right) \\ & \quad = P\left(A_{(1)} \leq x\right) = 1 - P\left(A_{(1)} > x\right) \\ & \quad = 1 - (1 - F(x))^m . \end{aligned} \tag{26}$$

Secondly, we note that the event $\left\{A_{(1)} \leq x,\ A_{(2)} > y\right\}$ can be partitioned into the m events $\{A_j \leq x,\ A_h > y, h \neq j\}$, $j = 1, \ldots, m$, and that

$$P\left(A_j \leq x,\ A_h > y, h \neq j\right) = F(x)\left(1 - F(y)\right)^{m-1} \tag{27}$$

holds. Combining this partitioning with (27) and (26) we arrive at (14).

Furthermore, we have

$$\begin{aligned} & P\left(A_{(1)} \leq \tilde{A}_{(1)},\ A_{(2)} \leq \tilde{A}_{(2)}\right) = E\left(P\left(A_{(1)} \leq \tilde{A}_{(1)},\ A_{(2)} \leq \tilde{A}_{(2)} \,|\, \tilde{A}_{(1)} \leq \tilde{A}_{(2)}\right)\right) \\ & \quad = E\left(1 - \left(1 - F(\tilde{A}_{(1)})\right)^m - mF(\tilde{A}_{(1)})\left(1 - F(\tilde{A}_{(2)})\right)^{m-1}\right) \\ & \quad = \int_{-\infty}^{\infty}\int_{x}^{\infty}\left[1 - (1 - F(x))^m - mF(x)\left(1 - F(y)\right)^{m-1}\right] \\ & \qquad\qquad\qquad m(m-1)\left(1 - F(y)\right)^{m-2} dF(y)\, dF(x) \\ & \quad = \int_0^1\int_u^1\left[1 - (1-u)^m - mu\,(1-v)^{m-1}\right] m(m-1)\,(1-v)^{m-2}\, dv\, du \\ & \quad = m(m-1)\int_0^1\left[\{1 - (1-u)^m\}\frac{1}{m-1}(1-u)^{m-1} - mu\frac{1}{2m-2}(1-u)^{2m-2}\right] du \\ & \quad = m\int_0^1\left[(1-u)^{m-1} - (1-u)^{2m-1} - \frac{1}{2}m(1-u)^{2m-2} + \frac{1}{2}m(1-u)^{2m-1}\right] du \\ & \quad = 1 - \frac{1}{2} - \frac{1}{2}\frac{m^2}{2m-1} + \frac{1}{4}m \\ & \quad = \frac{3m-2}{4(2m-1)}. \end{aligned} \tag{28}$$

Finally, as $(3m-2)/(4(2m-1))$ is increasing in m we obtain (16).

B Explanatory analyses of the data from Gillebaart et al. (2012)

Gillebaart et al. (2012), henceforth GFR, report six experiments (1A, 1B, 1C, 2A, 2B, 3). Experiment 1C has two groups of participants; all other experiments have three groups (Promotion, Control, Prevention). Each group is subjected to four conditions in which three Hebrew letters were exposed 0, 5, 15 or 40 times, subliminally, for 14 milliseconds. In each of the four Exposure conditions, each participant rates the attractiveness of three Greek letters and the average of the three item responses is the participant's Liking score in that condition.

B.1 Descriptives

Means and standard deviations of Liking scores are given in Table 14, plots of mean scores are given by Figures 5 and 6. The figures show consistent ordering of the Liking score means of the four Exposure conditions for the Prevention group in each of the six studies.

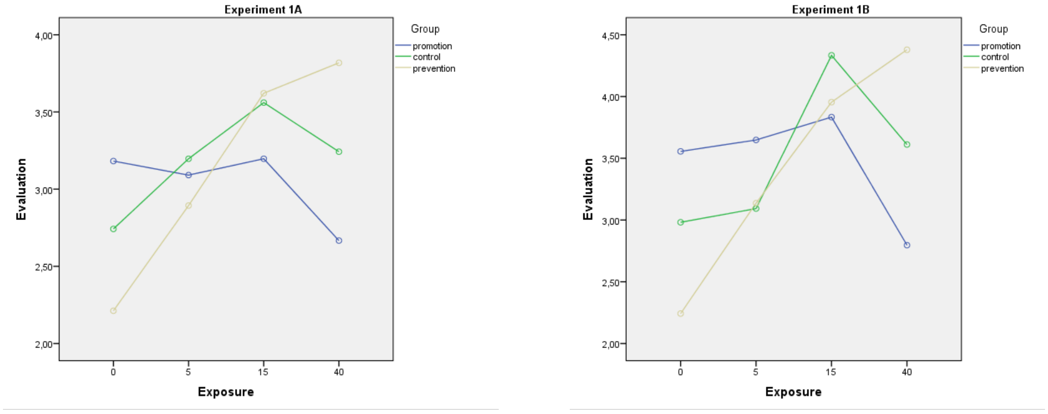


Figure 5. *Plots of group means of Experiments 1A and 1B.*

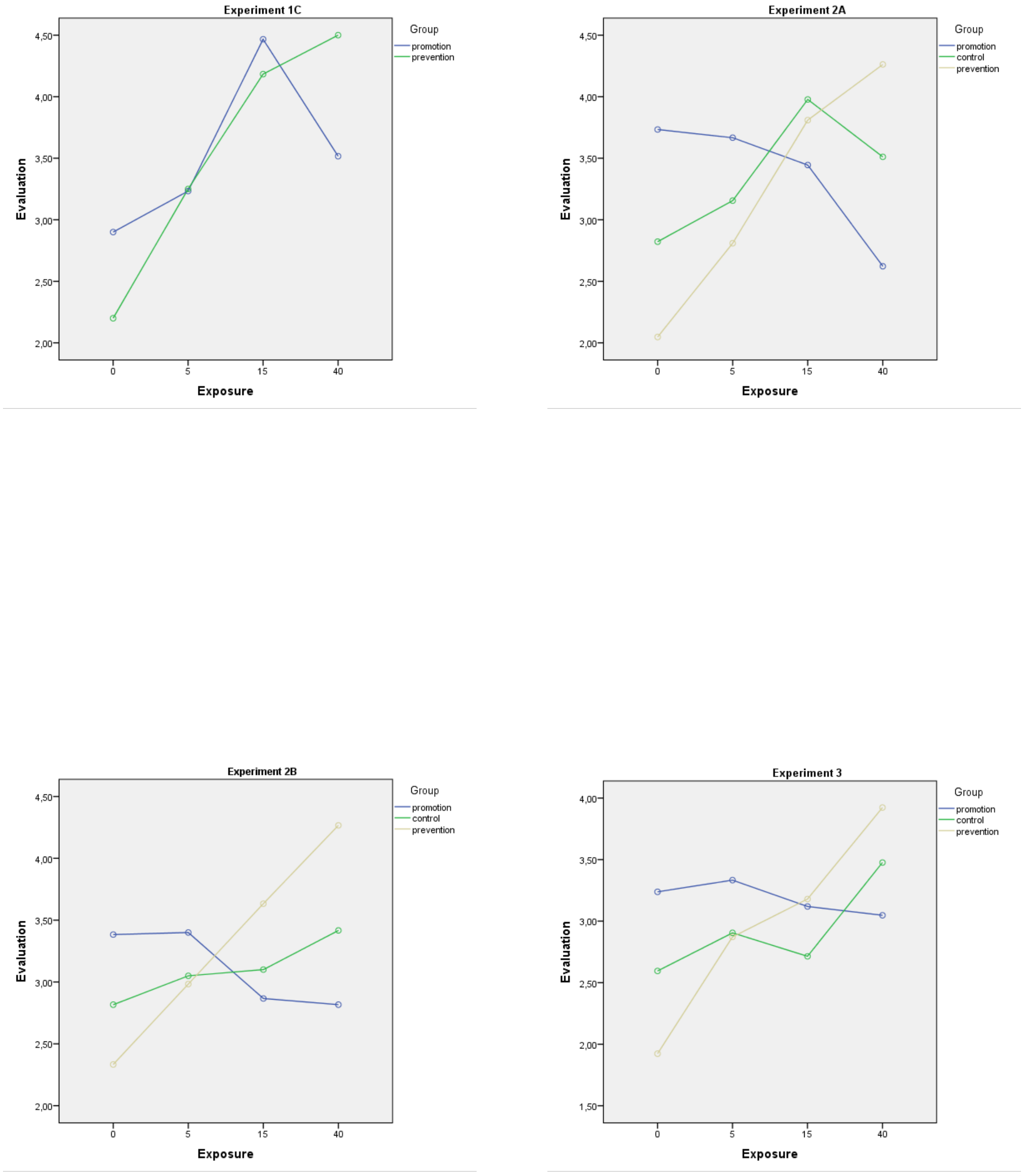


Figure 6. *Plots of group means of Experiments 1C, 2A, 2B, and 3.*

Table 14: *Means and standard deviations of observed Liking scores*

		Exposure 0	Exposure 5	Exposure 15	Exposure 40
	N	Mean SD	Mean SD	Mean SD	Mean SD
Experiment 1A					
Group Promotion	22	3.182 0.985	3.091 0.785	3.197 0.883	2.667 0.943
Group Control	22	2.742 0.748	3.197 0.648	3.561 0.567	3.242 1.200
Group Prevention	22	2.212 0.499	2.894 0.670	3.621 0.772	3.818 0.606
Experiment 1B					
Group Promotion	18	3.556 0.984	3.648 0.690	3.833 0.834	2.796 1.329
Group Control	18	2.981 0.889	3.093 0.782	4.333 0.922	3.611 0.629
Group Prevention	22	2.242 0.750	3.136 1.032	3.955 1.080	4.379 0.972
Experiment 1C					
Group Promotion	20	2.900 1.114	3.233 0.497	4.467 0.704	3.517 0.841
Group Prevention	20	2.200 0.661	3.250 0.601	4.183 1.172	4.500 1.084
Experiment 2A					
Group Promotion	15	3.733 1.267	3.667 0.398	3.444 0.965	2.622 1.030
Group Control	15	2.822 0.950	3.156 0.711	3.978 1.493	3.511 0.628
Group Prevention	14	2.048 0.885	2.810 0.748	3.810 1.019	4.262 0.786
Experiment 2B					
Group Promotion	20	3.383 0.913	3.400 0.777	2.867 1.005	2.817 1.147
Group Control	20	2.817 0.411	3.050 0.575	3.100 0.668	3.417 0.591
Group Prevention	20	2.333 0.865	2.983 0.841	3.633 0.601	4.267 0.896
Experiment 3					
Group Promotion	14	3.238 0.842	3.333 0.613	3.119 0.873	3.048 0.815
Group Control	14	2.595 1.006	2.905 0.497	2.714 0.702	3.476 1.160
Group Prevention	13	1.923 0.364	2.872 0.519	3.179 1.077	3.923 1.081

B.2 Statistical tests and effect sizes

Table 15 gives the results of the repeated measures ANOVA. These analyses have also been reported by GFR, but they gave incorrect partial η^2 values.

Table 15: *Repeated measures ANOVA results*

	F	df1	df2	probability	partial η^2	observed power
Experiment 1A						
Group	0.710	2	63	0.496	0.022	0.165
Exposure	10.936	3	189	0.000	0.148	0.999
Group × Exposure	7.524	6	189	0.000	0.193	1.000
Experiment 1B						
Group	0.143	2	55	0.867	0.005	0.071
Exposure	14.561	3	165	0.000	0.209	1.000
Group × Exposure	9.092	6	165	0.000	0.248	1.000
Experiment 1C						
Group	0.001	1	38	0.980	0.000	0.050
Exposure	38.663	3	114	0.000	0.504	1.000
Group × Exposure	7.803	3	114	0.000	0.170	0.987
Experiment 2A						
Group	0.451	2	41	0.640	0.022	0.118
Exposure	6.447	3	123	0.000	0.136	0.966
Group × Exposure	8.293	6	123	0.000	0.288	1.000
Experiment 2B						
Group	1.940	2	57	0.153	0.064	0.386
Exposure	6.440	3	171	0.000	0.102	0.967
Group × Exposure	9.511	6	171	0.000	0.250	1.000
Experiment 3						
Group	1.837	2	38	0.173	0.088	0.359
Exposure	7.507	3	114	0.000	0.165	0.984
Group × Exposure	4.157	6	114	0.001	0.180	0.972

Main effects of Exposure and interaction effects of Group × Exposure are highly significant. Post hoc calculations of statistical power for these effects yield values that are consistently higher than 0.96, which is conspicuous with sample sizes ranging from 14 to 22 participants per group. Relatedly, the effect sizes are extremely large, much larger than effect sizes that are commonly found in the social sciences in general, or in other studies of the mere exposure effect in particular (according to a meta-analysis of Bornstein (1989)). Partial eta-squared values are difficult to interpret (Levine and Hullett (2002)), but η^2 of 0.01, 0.06, and 0.14 are considered to represent small, medium, and large effect sizes according to Cohen (1988) and

Richardson (2011).

To get an idea of the size of the between and within group differences, Table 16 gives effect sizes (d) that express

- the differences between Groups Promotion and Prevention for the Exposure 0 and Exposure 40 conditions,
- the differences within Groups Promotion and Prevention between the Exposure 0 and Exposure 40 conditions,
- the size of the interaction effects, based on the differences between the within-group differences in Groups Promotion and Prevention.

Effect sizes and their variances are estimated using the formulas given by Lipsey and Wilson (2000) and Borenstein et al. (2009).

Table 16: *Effect sizes for differences between and within groups.*

Differences between Group Promotion and Group Prevention				
	Exposure 0		Exposure 40	
	Effect size d	Variance(d)	Effect size d	Variance(d)
Experiment 1A	1.242	0.108	-1.453	0.115
Experiment 1B	1.515	0.130	-1.374	0.125
Experiment 1C	0.764	0.107	-1.013	0.113
Experiment 2A	1.531	0.179	-1.779	0.193
Experiment 2B	1.181	0.117	-1.409	0.125
Experiment 3	2.000	0.223	-0.919	0.164
Differences between Exposure 0 and Exposure 40				
	Group Prom.		Group Prev.	
	Effect size d	Variance(d)	Effect size d	Variance(d)
Experiment 1A	0.534	0.098	-2.898	0.608
Experiment 1B	0.648	0.125	-2.452	0.301
Experiment 1C	-0.618	0.077	-2.524	0.330
Experiment 2A	0.963	0.224	-2.647	0.924
Experiment 2B	0.549	0.168	-2.196	0.465
Experiment 3	0.230	0.195	-2.600	0.875
Interaction effects (differences between within-group differences)				
	Effect size d	Variance(d)		
Experiment 1A	2.697	0.285		
Experiment 1B	2.849	0.280		
Experiment 1C	1.784	0.183		
Experiment 2A	3.282	0.544		
Experiment 2B	2.606	0.367		
Experiment 3	2.736	0.528		

The effect sizes are very large, especially within Group Prevention, with $|d|$ much larger than 2 in each of the six experiments. As a result,the interaction effect sizes are also very large. In the social sciences, effect sizes d of 0.2, 0.5, and 0.8 are considered to represent small, medium, and large effect sizes according to the rules of thumb from Cohen (1988).

The extremely large effect sizes are remarkable, especially in view of the subtle experimental differences between and within groups.

B.3 Correlations between repeated measures

When we calculated the effect sizes reported in Table 16, we unexpectedly found some negative correlations between Liking scores. Table 17 gives all correlations between Liking scores, for each group, for each experiment.

Table 17: *Correlations between Liking scores; single, double, and triple asterisks indicate* $p < 0.10, p < 0.05,$ *and* $p < 0.01,$ *respectively.*

Experiment 1A		Exposure 0	Exposure 5	Exposure 15	Exposure 40
Group Prom.	Exposure 0	1			
	Exposure 5	-0.194	1		
	Exposure 15	-0.067	-0.195	1	
	Exposure 40	0.057	**0.443	-0.102	1
Group Control	Exposure 0	1			
	Exposure 5	0.088	1		
	Exposure 15	0.232	-0.156	1	
	Exposure 40	0.297	**0.480	-0.085	1
Group Prev.	Exposure 0	1			
	Exposure 5	-0.056	1		
	Exposure 15	0.232	0.062	1	
	Exposure 40	-0.286	0.146	-0.166	1
Experiment 1B		Exposure 0	Exposure 5	Exposure 15	Exposure 40
Group Prom.	Exposure 0	1			
	Exposure 5	-0.109	1		
	Exposure 15	***-0.725	0.096	1	
	Exposure 40	0.072	0.124	-0.174	1
Group Control	Exposure 0	1			
	Exposure 5	***0.660	1		
	Exposure 15	***-0.638	**-0.580	1	
	Exposure 40	-0.283	-0.334	0.248	1
Group Prev.	Exposure 0	1			
	Exposure 5	**0.468	1		
	Exposure 15	0.177	0.001	1	
	Exposure 40	0.173	-0.049	0.183	1
Experiment 1C		Exposure 0	Exposure 5	Exposure 15	Exposure 40

Group Prom.	Exposure 0	1			
	Exposure 5	0.256	1		
	Exposure 15	0.137	0.241	1	
	Exposure 40	0.358	-0.024	0.055	1
Group Prev.	Exposure 0	1			
	Exposure 5	-0.088	1		
	Exposure 15	-0.253	**0.496	1	
	Exposure 40	0.212	-0.067	0.205	1
Experiment 2A		Exposure 0	Exposure 5	Exposure 15	Exposure 40
Group Prom.	Exposure 0	1			
	Exposure 5	0.314	1		
	Exposure 15	0.123	*0.475	1	
	Exposure 40	-0.150	0.232	0.029	1
Group Control	Exposure 0	1			
	Exposure 5	0.044	1		
	Exposure 15	0.126	-0.310	1	
	Exposure 40	-0.090	-0.333	-0.182	1
Group Prev.	Exposure 0	1			
	Exposure 5	-0.411	1		
	Exposure 15	0.143	**-0.545	1	
	Exposure 40	-0.437	0.324	-0.136	1
Experiment 2B		Exposure 0	Exposure 5	Exposure 15	Exposure 40
Group Prom.	Exposure 0	1			
	Exposure 5	0.094	1		
	Exposure 15	**-0.508	0.117	1	
	Exposure 40	**-0.460	0.119	0.409	1
Group Control	Exposure 0	1			
	Exposure 5	0.016	1		
	Exposure 15	0.006	0.047	1	
	Exposure 40	*0.379	0.022	0.037	1
Group Prev.	Exposure 0	1			
	Exposure 5	0.161	1		
	Exposure 15	*-0.438	*-0.394	1	
	Exposure 40	-0.362	-0.165	*0.387	1

Experiment 3		Exposure 0	Exposure 5	Exposure 15	Exposure 40
Group Prom.	Exposure 0	1			
	Exposure 5	0.000	1		
	Exposure 15	-0.100	-0.144	1	
	Exposure 40	-0.329	0.188	0.268	1
Group Control	Exposure 0	1			
	Exposure 5	-0.066	1		
	Exposure 15	0.102	0.014	1	
	Exposure 40	-0.013	-0.316	- 0.376	1
Group Prev.	Exposure 0	1			
	Exposure 5	-0.204	1		
	Exposure 15	0.274	-0.121	1	
	Exposure 40	-0.299	0.179	-0.083	1

One might expect positive correlations between repeated measures, and higher correlations between exposure frequencies that are more similar. However, no such pattern exists and the overall average of all correlations equals 0.008. Only 17 out of 102 correlations are significant at a two sided level of significance of 10% (9 positive, 8 negative).

Table 18 gives the results of Likelihood Ratio tests of independence between Liking scores from different Exposure conditions.

Table 18: *Likelihood ratio tests of independence of Liking scores*

		χ^2_6	probability
Experiment 1A	Group Prom.	7.363	0.289
	Group Cont.	10.021	0.124
	Group Prev.	4.017	0.674
Experiment 1B	Group Prom.	14.664	0.023
	Group Cont.	23.432	0.001
	Group Prev.	7.966	0.241
Experiment 1C	Group Prom.	5.768	0.450
	Group Prev.	10.485	0.106
Experiment 2A	Group Prom.	7.300	0.294
	Group Cont.	5.250	0.512
	Group Prev.	11.088	0.086
Experiment 2B	Group Prom.	13.173	0.040
	Group Cont.	3.179	0.786
	Group Prev.	12.013	0.062
Experiment 3	Group Prom.	3.816	0.702
	Group Cont.	4.019	0.674
	Group Prev.	3.055	0.802

When testing at a 5% level of significance, the hypothesis of independence can only be rejected in three of the 17 groups: Groups Promotion and Control in Experiment 1B and Group Promotion in Experiment 2B .

Of course, with repeated measures we would not expect independence, but it is striking that the hypothesis of independence is only rejected in groups that show negative correlations between repeated measures. We might consider the significant results as chance results, which led us to investigate the reliability of the Liking scores.

B.4 Reliability of Liking scores

For each Exposure condition, each participant's Liking score is calculated by averaging the responses to three out of twelve Hebrew letters. Across the four Exposure conditions, each participant responded to the same 12 Hebrew letters, but the distribution of the letters across conditions was different for different participants. We therefore used generalizability coefficients to estimate the reliability of the Liking score; Cardinet et al. (1981).

In the social sciences, the usual rule of thumb is that reliability should be larger than 0.7 for measurements to be considered sufficiently reliable. However, in the GFR experiments only 8 out of the 68 generalizibility coefficients within the Group × Exposure conditions are higher than 0.7. Remarkably, 24 out of 68 generalizibility coefficients are zero.

The last column of Table 19 gives generalizability coefficients that are calculated across Exposure conditions, on the basis of 12 items. These calculations yield 8 zero values and no values above 0.7.

Low reliabilities in Table 19 already indicate that items may not be correlated. Indeed, the overall average inter-item correlation is 0.030, and only 12% of all inter-item correlations (135 out of 1122) are significant at a two-sided 10% level of significance. Apparently, there are no respondent characteristics (or response styles or tendencies) that affect the item responses. If this were true, then items within conditions should be independent. Table 20 gives the results of likelihood ratio tests of independence of item responses within Exposure conditions.

The hypothesis of independence is rejected 24/68 times at a 5% level of significance, and 15/68 times at 1% level of significance. This is much more often than expected under the null hypothesis. So we conclude that the items are generally not independent, which is strange in combination with the overall average of 0.030. There is large variation in inter-item correlations, with about as many positive correlations as negative correlations. The variation in correlations is unexpected, seeing that in the GFR studies all items have similar content and should have equivalent characteristics. In the next section we check the interchangeability of items.

Table 19: *Generalizibility coefficients; coefficients are calculated on the basis of REML estimates of variance components*

	Exposure 0 3 items	Exposure 5 3 items	Exposure 15 3 items	Exposure 40 3 items	Across 12 items
Experiment 1A					
Group Prom.	0.683	0.500	0.151	0.619	0.336
Group Control	0.306	0.422	0.000	0.846	0.600
Group Prev.	0.000	0.000	0.000	0.000	0.000
Experiment 1B					
Group Prom.	0.445	0.000	0.000	0.504	0.000
Group Control	0.289	0.000	0.000	0.000	0.000
Group Prev.	0.291	0.554	0.122	0.150	0.287
Experiment 1C					
Group Prom.	0.713	0.235	0.000	0.544	0.462
Group Prev.	0.416	0.200	0.620	0.702	0.310
Experiment 2A					
Group Prom.	0.728	0.000	0.213	0.225	0.313
Group Control	0.524	0.000	0.884	0.000	0.000
Group Prev.	0.463	0.007	0.748	0.000	0.000
Experiment 2B					
Group Prom.	0.474	0.000	0.692	0.557	0.239
Group Control	0.000	0.000	0.289	0.000	0.000
Group Prev.	0.570	0.544	0.000	0.335	0.000
Experiment 3					
Group Prom.	0.563	0.000	0.459	0.462	0.075
Group Control	0.663	0.000	0.000	0.664	0.000
Group Prev.	0.000	0.160	0.823	0.883	0.132

B.5 Interchangeability of items

According to GFR: "Target stimuli consisted of 12 Hebrew letters, pretested as valence neutral and presented 0, 5, 15, or 40 times within participants in the mere exposure phase of the experiment. Assignment of the letters to the exposure frequencies was fully counterbalanced over all participants, and order of stimuli was randomized for each participant" (p. 701). We therefore expect item means, variances, and covariances to be invariant within Group × Exposure conditions.

We tested the invariance of item means, variances, covariances through likelihood ratio tests, for each Group × Exposure condition, by comparing the fit of a model in which item means, variances, and covariances are free to be estimated with a model in which means are constrained to be equal to each other, variances are constrained to be equal to each other, and

Table 20: *Likelihood ratio tests of independence of item responses*

	Exposure 0		Exposure 5		Exposure 15		Exposure 40	
	χ^2_3	Prob.	χ^2_3	Prob.	χ^2_3	Prob.	χ^2_3	Prob.
Experiment 1A								
Group Prom.	10.896	0.012	6.769	0.080	0.999	0.801	9.686	0.021
Group Control	2.620	0.454	2.839	0.417	4.811	0.186	32.650	0.000
Group Prev.	2.468	0.481	7.806	0.050	0.560	0.906	1.164	0.762
Experiment 1B								
Group Prom.	7.006	0.072	3.296	0.348	0.291	0.962	3.690	0.297
Group Control	4.639	0.200	6.124	0.106	2.427	0.489	4.600	0.204
Group Prev.	5.597	0.133	9.354	0.025	5.576	0.134	3.311	0.346
Experiment 1C								
Group Prom.	12.608	0.006	4.954	0.175	1.550	0.671	8.192	0.042
Group Prev.	4.984	0.173	0.911	0.823	8.879	0.031	14.626	0.002
Experiment 2A								
Group Prom.	14.563	0.002	14.597	0.002	2.044	0.563	4.062	0.255
Group Control	4.416	0.220	5.133	0.162	27.320	0.000	5.134	0.162
Group Prev.	24.285	0.000	12.544	0.006	17.456	0.001	4.082	0.253
Experiment 2B								
Group Prom.	14.263	0.003	0.262	0.967	12.030	0.007	8.425	0.038
Group Control	9.201	0.027	1.591	0.661	7.626	0.054	3.279	0.351
Group Prev.	7.914	0.048	6.278	0.099	1.624	0.654	4.501	0.212
Experiment 3								
Group Prom.	5.480	0.140	1.739	0.628	16.895	0.001	5.297	0.151
Group Control	15.462	0.001	6.473	0.091	4.608	0.203	11.024	0.012
Group Prev.	1.777	0.620	5.042	0.169	17.999	0.000	24.186	0.000

covariances are constrained to be equal to each other. Table 21 gives the results.

Hypotheses of invariance are rejected much more often than expected if the items indeed are interchangeable and have equal content. For example, at the 5% level of significance, the global test of equal means, variances, and covariances has to be rejected 7/17, 9/17, 9/17, and 2/17 times for the four Exposure conditions, and 27/68 times overall (under the null hypothesis of invariance, with 17 tests, the probability of finding 7 or more significant results at 5% is 9.7×10^{-6}, and the probability of finding 9 or more significant results at 5% is 3.3×10^{-8}).

Table 21: *Likelihood ratio tests of equality of item means, variances, and covariances*

	Exposure 0		Exposure 5		Exposure 15		Exposure 40	
	χ^2_6	Prob.	χ^2_6	Prob.	χ^2_6	Prob.	χ^2_6	Prob.
Experiment 1A								
Group Prom.	4.571	0.600	8.733	0.189	5.003	0.543	7.301	0.294
Group Control	5.896	0.435	2.887	0.823	5.412	0.492	9.612	0.142
Group Prev.	27.529	0.000	11.219	0.082	11.404	0.077	3.345	0.764
Experiment 1B								
Group Prom.	9.070	0.170	4.485	0.611	16.921	0.010	1.750	0.941
Group Control	18.738	0.005	14.194	0.028	14.709	0.023	10.517	0.105
Group Prev.	17.763	0.007	16.653	0.011	30.512	0.000	20.060	0.003
Experiment 1C								
Group Prom.	3.138	0.791	12.727	0.048	17.485	0.008	16.430	0.012
Group Prev.	10.692	0.098	13.608	0.034	4.698	0.583	6.677	0.352
Experiment 2A								
Group Prom.	10.729	0.097	10.518	0.104	13.604	0.034	9.130	0.166
Group Control	10.865	0.093	20.642	0.002	5.817	0.444	9.442	0.150
Group Prev.	24.646	0.000	27.671	0.000	10.083	0.121	2.668	0.849
Experiment 2B								
Group Prom.	19.862	0.003	16.486	0.011	4.883	0.559	7.863	0.248
Group Control	10.018	0.124	10.689	0.098	13.039	0.042	5.172	0.522
Group Prev.	7.432	0.283	12.041	0.061	17.520	0.008	12.351	0.055
Experiment 3								
Group Prom.	6.339	0.386	24.139	0.000	30.235	0.000	10.741	0.097
Group Control	12.947	0.044	20.299	0.002	19.561	0.003	6.020	0.421
Group Prev.	26.577	0.000	7.208	0.302	9.994	0.125	5.305	0.505

B.6 Conclusion

Apparently, the items are not valence neutral, nor counterbalanced, nor randomized across Exposure conditions. This may be an honest mistake, but the GFR conclusions that variance between Liking scores across Group × Exposure conditions is explained by variations in regulatory focus and novelty categorization hinges on measurement equivalence across Group × Exposure conditions. However, if there is no measurement invariance within Group × Exposure conditions, then there cannot be measurement invariance across conditions either.

Moreover, GFR assume that regulatory focus and novelty categorization affect Liking. However, as item responses are not correlated and generalizability coefficients indicate that the Liking scores are random error rather than representing anything structural, effects of regulatory focus or novelty categorization on the Liking scores cannot be interpreted as influences of the attractiveness of Hebrew letters.

References

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009). *Introduction to Meta-Analysis.* Wiley, New York.

Bornstein, R.F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968-1987. *Psychological Bulletin*, **106**, 265–289.

Cardinet, J., Tourneur, Y, and Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, **18**, 183-204.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Routledge Academic, New York.

Fairbanks, D.J. and Rytting, B. (2001). Mendelian controversies: a botanical and historical review. *American Journal of Botany* **88**, 737–752.

Gillebaart, M., Förster, J. and Rotteveel, M. (2012). Mere Exposure Revisited: The Influence of Growth Versus Security Cues on Evaluations of Novel and Familiar Stimuli. *Journal of Experimental Psychology: General* **141**, 699-714.

Gillebaart, M., Förster, J., Rotteveel, M. and Jehle, A.C.M. (2013). Unraveling effects of novelty on creativity. *Creativity Research Journal*, **25**, 280–285.

Klaassen, C.A.J. (2015). Evidential Value in ANOVA-Regression Results in Scientific Integrity Studies. *arXiv:1405.4540v2.*

Levine, T.R. and Hullett, C.R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Human Communication Research*, **28**, 612-625.

Lipsey, M.W. and Wilson, D.B. (2000). Practical Meta-Analysis. *Applied Social Research Methods*, **49**. Sage Publications, London.

Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn (Abhandlungen)* **4**, 3-47.

Peeters, C.F.W., Klaassen, C.A.J. and Van de Wiel, M.A. (2015). Evaluating the Scientific Veracity of Publications by dr. Jens Förster. *Report.*

Richardson, J.T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, **6**, 135–147.